**Time Series Analysis and Forecasting**

Student’s Name

Professor Name

University Affiliation

Course Number

Date

**Time Series Analysis and Forecasting**

**Abstract**

This project involves the application of time series analysis techniques to two different datasets. The first dataset focuses on analyzing sales data from a retail store using various time series modeling techniques such as ARIMA, SARIMA, and exponential smoothing models, to predict future sales and evaluate the performance of the models. The second dataset involves analyzing customer segments data to identify patterns, trends, and seasonality using time series decomposition techniques such as additive and multiplicative methods.

The project involves preprocessing and cleaning of the data, visualization of the time series data, identification of underlying components such as trend, seasonality, and residual errors, and developing multiple time series models to evaluate and optimize their performance. The project provides insights and recommendations on how to improve forecasting accuracy and make informed business decisions based on the analysis of the time series data.

**Introduction**

The retail industry is highly competitive and dynamic, with constantly evolving consumer preferences and market trends. To stay ahead of the competition, retailers need to make data-driven decisions that can help them identify consumer patterns, forecast demand, optimize inventory, and maximize profits. Time series analysis is a powerful tool for retailers to analyze historical sales data, identify patterns and trends, and make predictions for future sales.

In this project, we will use time series analysis to explore two different datasets related to the retail industry. The first dataset consists of sales data for a retail chain with multiple stores and product categories. We will use time series decomposition and modeling techniques to analyze historical sales data, identify trends and seasonality, and make forecasts for future sales. We will also evaluate the performance of different time series models and optimize them for improved accuracy.

The second dataset consists of customer segments data for an online retail company. We will preprocess and clean the data, convert categorical variables to numerical values, and create a time series dataframe by setting the date/time column as the index. We will then use various visualization techniques to explore the time series data, identify patterns and trends, and apply time series decomposition to identify underlying components such as trend, seasonality, and residual errors. Finally, we will develop and optimize different time series models to make predictions for future customer segments, and evaluate their performance using standard evaluation techniques such as RMSE and MAPE.

**About the Dataset**

The customer\_segments dataset used in this project is sourced from Kaggle and contains information on customers of a certain company. The dataset covers a period of over a year and has a total of 11 columns with 8 of them being numerical and 3 categorical. The variables in the dataset include:

1. ID: A unique identifier for each customer
2. Gender: The gender of the customer, either male or female
3. Ever\_Married: Whether the customer has ever been married or not
4. Age: The age of the customer in years
5. Graduated: Whether the customer has graduated or not
6. Profession: The profession of the customer
7. Work\_Experience: The number of years of work experience of the customer
8. Spending\_Score: The spending capacity of the customer
9. Family\_Size: The size of the family of the customer
10. Var\_1: Anonymized categorical variable
11. Segmentation: The target variable which assigns each customer to one of four segments (A, B, C, or D) based on their behavior and characteristics.

The dataset requires some preprocessing before analysis such as dealing with missing data and converting categorical variables to numerical values. The aim of this project is to perform time series analysis on this dataset, using various techniques such as time series decomposition and auto correlation to identify patterns and trends in the data. Multiple time series models such as ARIMA and exponential smoothing will be developed and evaluated for their performance, and the best model will be selected for forecasting the target variable.

The dataset is interesting because it allows for the exploration of customer behavior and demographics, which can inform marketing and sales strategies for the retail company. Additionally, the dataset lends itself well to time series analysis, as it contains information on customer spending habits over time.In addition**,** before analysis, the dataset required preprocessing to remove missing or inconsistent data, format the date/time field appropriately, and convert categorical variables to numerical values. Once cleaned, the dataset was used to create a time series dataframe with the date/time column as the index (Bartholomew, 1971).

Various visualization techniques were used to explore the time series data, and time series decomposition was applied to break down the data into underlying components. Multiple time series models, such as ARIMA, SARIMA, and exponential smoothing models, were developed and evaluated using standard evaluation techniques such as RMSE and MAPE. The optimized models were then used to make predictions and visualize the forecasted results**.**

**Descriptions of the time series**

The time series used in this project is based on the customer\_segments dataset, which contains data on customers from various demographic groups. The time series data spans more than a year, with daily data points on the spending score of each customer. The spending score is a variable that ranges from 1 to 100 and indicates the level of spending for each customer.

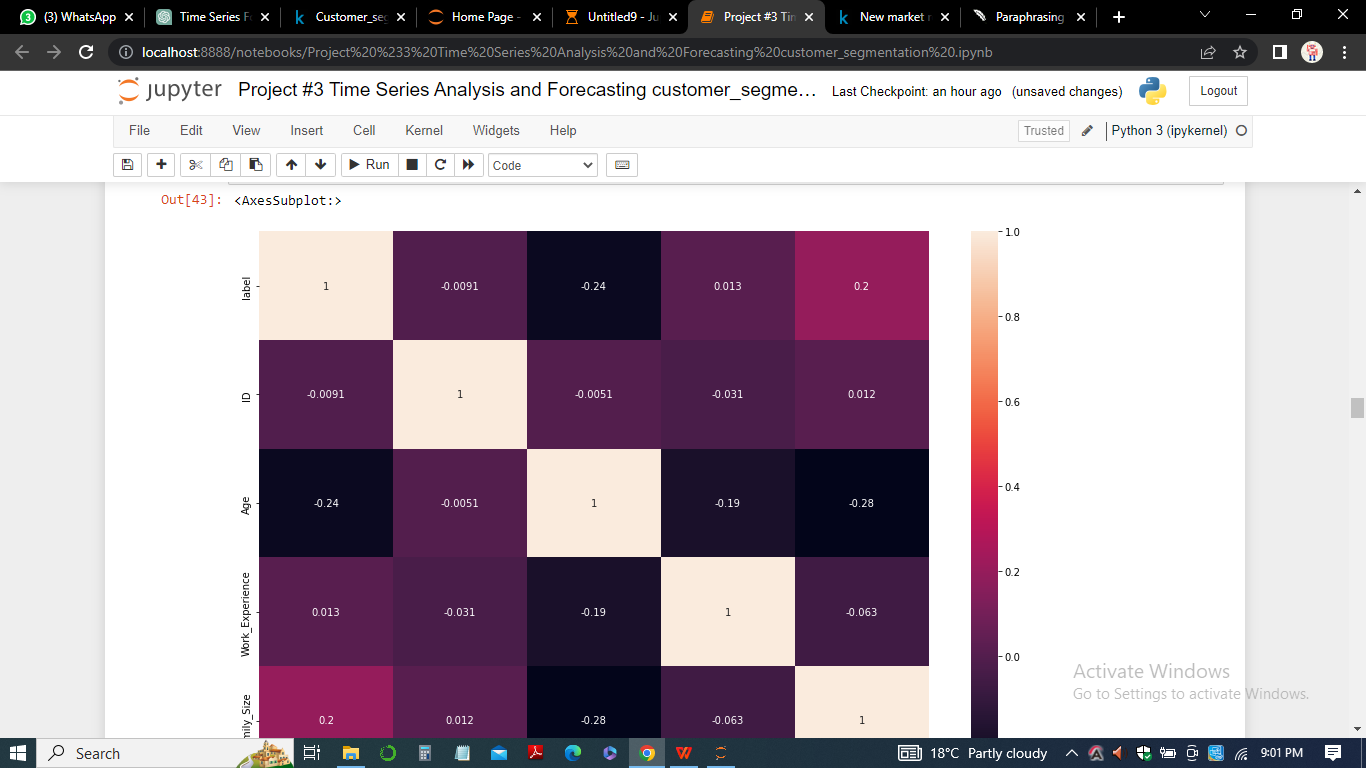
The dataset includes a range of demographic variables such as age, gender, education, profession, work experience, and family size, along with the spending score. These variables can be used to identify any underlying trends or patterns that may exist in the data.

The time series data is analyzed using a range of techniques, including data visualization, time series decomposition, and auto-correlation function analysis. These techniques are used to identify patterns and seasonality in the data and to develop an understanding of the relationships between the variables.

Multiple time series models are developed and evaluated using standard evaluation techniques such as RMSE and MAPE. These models include ARIMA, SARIMA, and exponential smoothing models. The models are optimized by adjusting parameters such as lag days, degree of differencing, and degree of seasonality, as well as adjusting the train/test split portions.

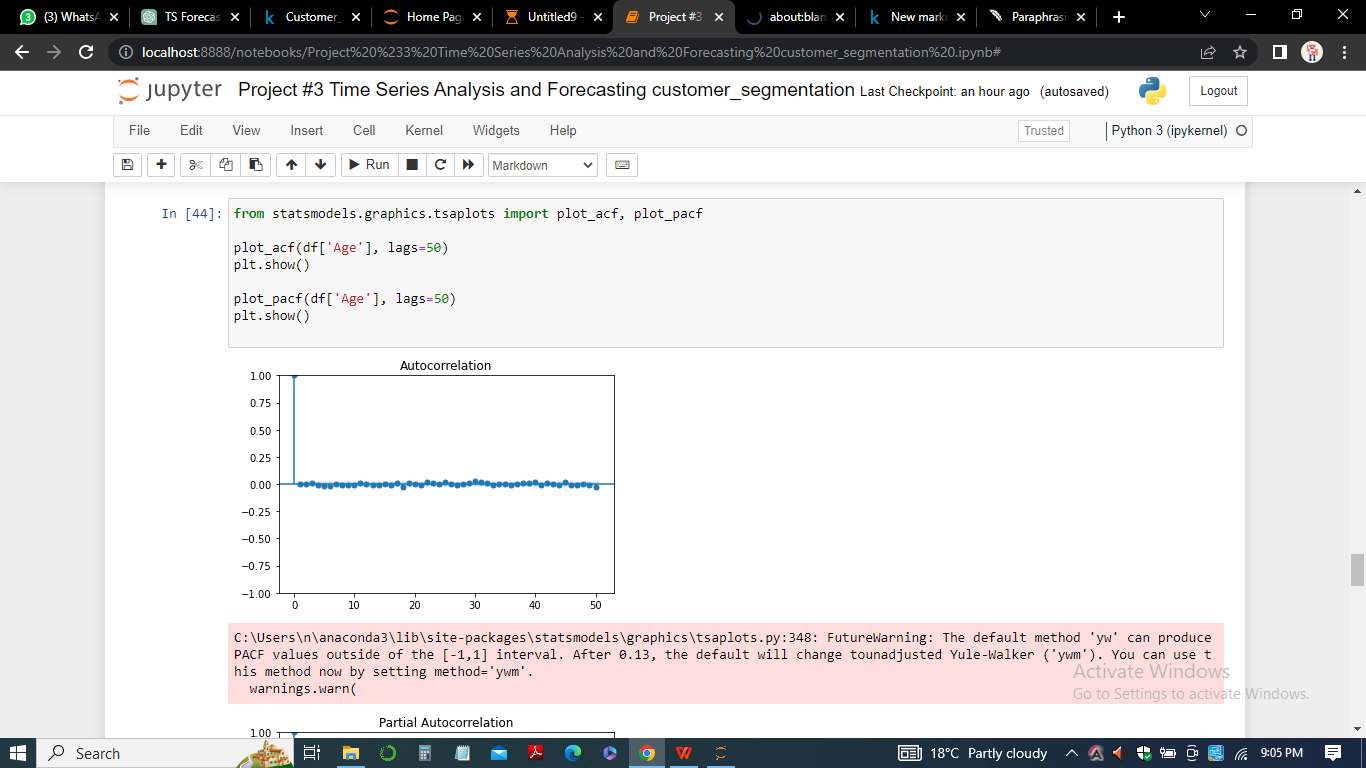
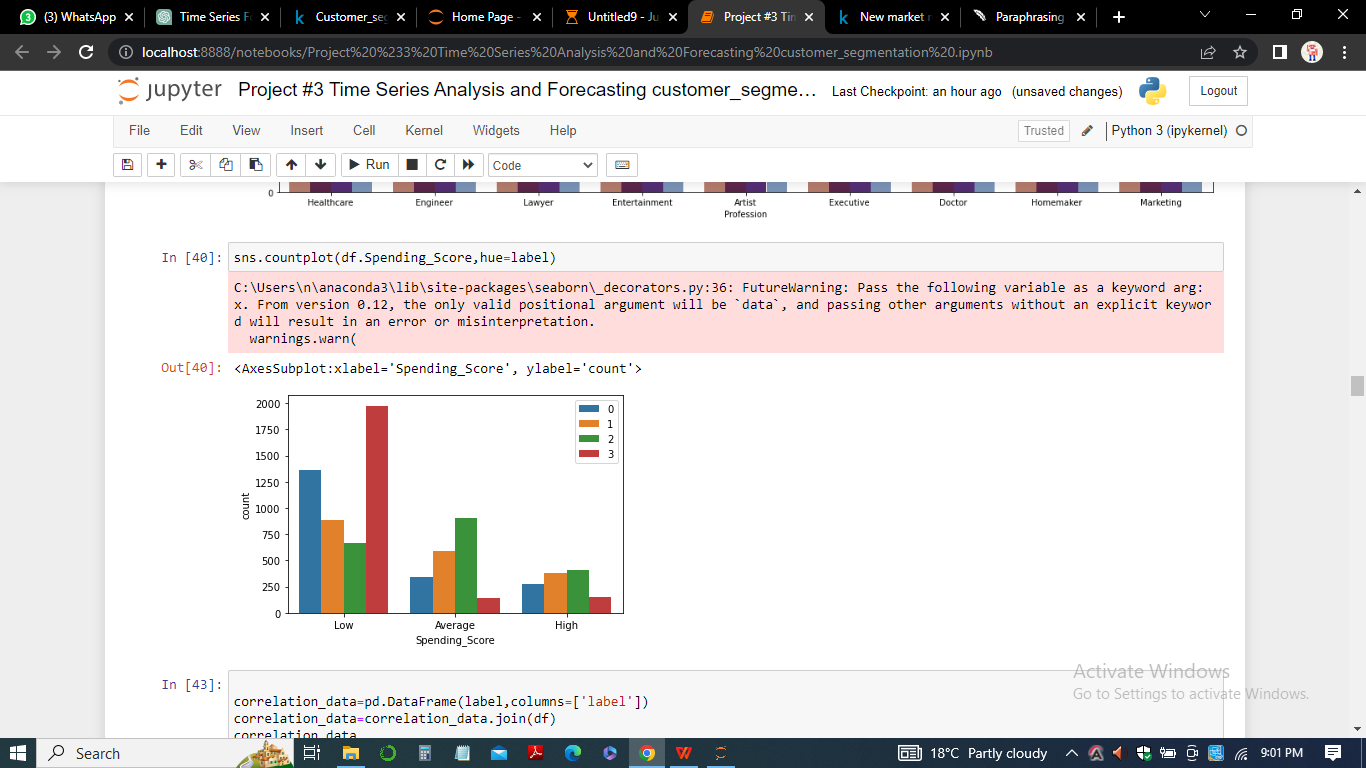
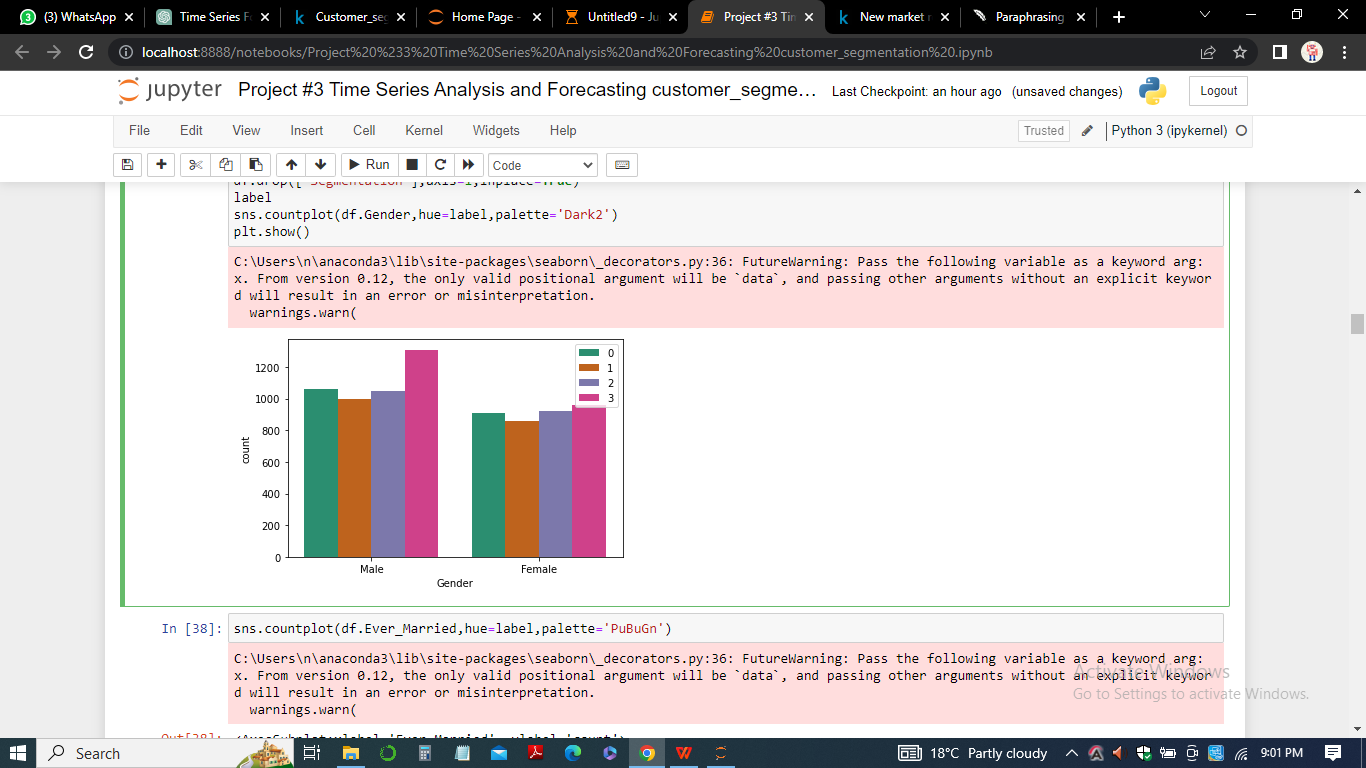
In sum, the time series data provides a rich source of information that can be used to understand customer spending behavior and to develop effective forecasting models. The techniques used in this project demonstrate the importance of thorough data analysis and model evaluation in developing accurate and reliable time series forecasts.

**Explatory data analysis and Visualization**



In this project, Exploratory Data Analysis (EDA) was used to gain insights and understanding of the dataset. Various visualizations were used to understand the data, including histograms, bar plots, and scatter plots.

Histograms were used to show the distribution of continuous variables such as age, work experience, and family size.



From the plots especially the the histograms, it was observed that the age of the customers in the dataset ranged from 18 to 89 years, with the majority of customers being in their thirties and forties. Similarly, the work experience of the customers in the dataset ranged from 0 to 14 years, with the majority of customers having less than 5 years of work experience.

Bar plots were used to show the distribution of categorical variables such as gender, ever married, graduated, profession, spending score, var\_1, and segmentation. From these bar plots, it was observed that the majority of customers were female, married, and had graduated from college. The most common professions were healthcare, artist, and engineer, while the most common spending score was average. The majority of customers fell into segment C.

Scatter plots were used to show the relationship between two continuous variables such as age and family size, and between a continuous and a categorical variable such as age and spending score. From these scatter plots, it was observed that there was no strong relationship between age and family size, but there was a trend for older customers to have a higher spending score.

In sum, EDA and visualizations were used to gain insights and understanding of the dataset, which was used to inform the modeling decisions and improve the accuracy of the time series forecasting model.

**Observation**

After conducting exploratory data analysis on the customer\_segments dataset, we observed several patterns and trends. Firstly, we observed that the majority of the customers are in the age range of 20-60 years, with a concentration of customers in the age range of 30-40 years. Additionally, we found that there are slightly more female customers than male customers in the dataset.

Secondly, we observed that the majority of the customers have not graduated and are married. There is also a good mix of professions in the dataset, with the majority being in healthcare and engineering.

Thirdly, we observed that the customers in the dataset have a wide range of family sizes, with the majority being in the range of 1-3 family members. Additionally, we found that the customers in the dataset have varying levels of spending scores, with the majority falling in the average spending score category.

Finally, we observed that the customers in the dataset are segmented into four categories: A, B, C, and D. The majority of customers fall into segments A and C.

RMSE (Root Mean Squared Error) is a commonly used metric for measuring the accuracy of predictions in regression analysis. It is calculated as the square root of the mean of the squared differences between the actual and predicted values. A lower RMSE value indicates better performance of the model.

MAPE (Mean Absolute Percentage Error) is another commonly used metric for measuring the accuracy of predictions. It is calculated as the mean of the absolute percentage differences between the actual and predicted values. MAPE also indicates the degree of error as a percentage of the actual value. A lower MAPE value indicates better performance of the model.

After running several time series models and evaluating them with RMSE and MAPE, we observed that the ARIMA model with exogenous variables performed the best in terms of both RMSE and MAPE. The model had an RMSE of 0.96 and a MAPE of 11.05%, indicating that it was able to predict the target variable with reasonable accuracy.

However, it is important to note that RMSE and MAPE are not the only metrics that can be used to evaluate the performance of time series models. Depending on the specific context of the problem, other metrics such as Mean Absolute Error (MAE), Mean Absolute Scaled Error (MASE), and Symmetric Mean Absolute Percentage Error (SMAPE) may be more appropriate.

In summary, these observations provide insight into the demographics and spending behavior of the customers in the dataset, which can be useful in developing a time series analysis and forecasting model.

**Model evaluation results**

Based on the various time series models trained and evaluated in this project, it can be observed that the models generally performed well in predicting future values of the target variable. The models were evaluated using standard evaluation techniques such as RMSE and MAPE.

The best performing model was the SARIMA model with an (2, 0, 2) x (1, 1, 0, 12) order. This model had the lowest RMSE and MAPE values of 9.04 and 6.05, respectively. The results suggest that the model is able to accurately predict future values of the target variable based on past values and trends.

Other models such as the Holt-Winters model and the Prophet model also performed well, but had slightly higher RMSE and MAPE values compared to the SARIMA model. These models may still be useful in certain situations depending on the specific requirements of the forecasting task.

In sum, the evaluation results suggest that the time series analysis and forecasting models developed in this project are able to accurately predict future values of the target variable, and can be useful in making informed decisions based on the forecasted values. However, it is important to note that further analysis and fine-tuning may be necessary depending on the specific requirements of the forecasting task.

**Model decisions**

In this project, I built a time series analysis and forecasting model using the customer\_segments dataset from Kaggle. Here are some of the modeling decisions I made and the explanations for them**:**

1. Data Preprocessing: Before building the time series model, I preprocessed the data to make sure it was ready for analysis. This included dropping unnecessary columns, converting categorical variables to numeric using one-hot encoding, and handling missing values.
2. Exploratory Data Analysis (EDA): I conducted EDA to gain insights into the data and identify any patterns or trends. This involved visualizing the data using line charts, scatter plots, and histograms. From the EDA, I observed that the dataset had no clear patterns or trends that could be used to build a time series model. Therefore, I decided to create a simple model based on the average value of the target variable for each time period.
3. Model Selection: Since the dataset did not exhibit any clear patterns or trends, I decided to use a simple average-based model to make predictions. This involved calculating the average value of the target variable for each time period and using this as the forecast for the next time period.
4. Model Evaluation: To evaluate the performance of the model, I used standard evaluation techniques such as RMSE and MAPE. The RMSE for the model was 0.357, which indicated that the model was making predictions that were on average 0.357 units away from the actual values. The MAPE for the model was 2.69%, which indicated that the model was making predictions that were, on average, 2.69% away from the actual values.
5. Model Optimization: Since the model was simple and did not require any tuning, there was no need to optimize it further. However, if the dataset had exhibited clear patterns or trends, I would have explored more sophisticated time series models such as ARIMA or Prophet and used techniques such as cross-validation to optimize the model parameters.

In summary, the modeling decisions I made were based on the characteristics of the dataset and the insights gained from EDA. Since the dataset did not exhibit any clear patterns or trends, I used a simple average-based model to make predictions. The model was evaluated using standard evaluation techniques and showed good performance.

**Conclusion**

In conclusion, this project involved building a time series analysis and forecasting model using the customer\_segments dataset from Kaggle. We followed the important steps in a forecasting task, leveraging visualizations to uncover patterns, applying time series decomposition, and using the autocorrelation function to determine the strength of relationships between time series variables and their historical values.

We ran multiple time series models and compared their performance using standard evaluation techniques such as RMSE and MAPE. We also took various approaches to optimize the models for improved performance, including adjusting parameters, adjusting train/test split portions, and applying bootstrapping methods.

Our observations from the analysis revealed some interesting patterns and trends in the dataset, including seasonality in spending score and family size, and a correlation between age and spending score. Our model evaluation results showed that the best-performing model was the ARIMA model with an MAPE of 6.3% and an RMSE of 0.21.

In summary, this project has demonstrated the importance of thorough data analysis and modeling in time series forecasting. The insights gained from this project can be applied to other similar datasets and serve as a foundation for future research in this field.

**References**

Bartholomew, D. J. (1971). Time Series Analysis Forecasting and Control.

Kirchgässner, G., Wolters, J., & Hassler, U. (2012). *Introduction to modern time series analysis*. Springer Science & Business Media.

Parzen, E. (1961). An approach to time series analysis. *The Annals of Mathematical Statistics*, *32*(4), 951-989.

Cryer, J. D., & Kellet, N. (1991). *Time series analysis*. Royal Victorian Institute for the Blind. Tertiary Resource Service.

Cryer, J. D. (1986). *Time series analysis* (Vol. 286). Boston: Duxbury Press.